Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi / Research Article

Makine Öğrenmesi Yöntemleri İle Eğitim Başarısının Tahmini Modeli

Prediction Model Of Educational Success With Machine Learning Method

Deniz Zilyas[1*], Atınç Yılmaz[2]

[1] İstanbul Beykent Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği A.B.D., kilitogludeniz@gmail.com ORCID: https://orcid.org/0000-0002-9454-9410

[2] İstanbul Beykent Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, atincyilmaz@beykent.edu.trORCID: https://orcid.org/0000-0003-0038-7519

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 3 Temmuz 2023
Revizyon 2 Ağustos 2023
Kabul 25 Eylül 2023
Online 30 Eylül 2023

Anahtar Kelimeler:

Makine Öğrenimi,
Tahmin,
Regresyon,
Eğitim Başarısı

ÖZ

Günümüzde makine öğrenmesi yöntemleri etkin bir biçimde kullanılarak pek çok alanda yüksek performanslar ve etkili sonuçlar göstermektedir. Bu nedenle makine öğrenmesi algoritmalarının uygulanması, çeşitli sektörlerde son yıllarda daha da yaygınlaşmıştır. Makine öğrenmesi modellerinden elde edilebilecek çıkarımlar ile birçok sorun öngörülüp çözüme ulaştırılabilir. Çalışmadaki amaç, ortaokul öğrencileri ile yapılan anket üzerinden elde edilen veriler kullanılarak; eğitim başarı tahminini yapacak bir makine öğrenmesi modeli ortaya koymak ve öğrenciyi olumsuz etkileyebilecek faktörleri belirlemektir. Anket soruları, öğrencinin başarısına tesir edebilecek etkenler literatürde araştırılarak oluşturulmuştur. Çalışma kapsamında, çeşitli ortaokullarda eğitim gören 519 farklı öğrenciden kişisel verilerin korunması kanunu kapsamında 13 sorudan oluşan anket aracılığıyla veri toplanmıştır. Bu veriler hiçbir kurumla paylaşılmamış olup, gizlilik korunmuştur. Veri seti ön işleme ve görselleştirme işlemlerinden sonra modelleme için, K-En Yakın Komşu (K-NN) , Rastgele Ormanlar (RO) , Lineer Regresyon, Bagged Trees Regression (BTR - Torbalanmış Ağaçlar), Gradient Boosting Regressor (GBM - Gradyen Arttırıcı Regresyon) ve Karar Ağaçları (KA) algoritmaları kullanılmıştır. Çalışmada, veri ön işleme adımları gerçekleştirildikten sonra makine öğrenmesi yöntemlerinin kullanımı ile oluşturulan model ile öğrencinin Türkçe notu üzerinden eğitim başarısının tahmini yapılmıştır. Çalışmada, ders seçiminin belirlenmesi, ana dilin Türkçe olması ve eğitim hayatından itibaren her dönem Türkçe dersi ile karşılaşılmasından dolayı Türkçe dersi bağımlı değişken olarak seçilmiştir. Çalışma neticesinde, rastgele orman yöntemi 0.88 doğruluk oranı ve 0.98 R-Kare değeri ile en etkin sonuçlar veren yöntem olmuştur.Öğrencinin eğitim durumunu etkileyen en önemli faktörler Türkçe notuna bağlı olarak aralarındaki korelasyon ile aile geliri ve ders çalışma saati olarak bulunmuştur.

ARTICLE INFO

Article history:

Received 3 July 2023
Received in revised from 2 August 2023
Accepted 25 September 2023
Available online 30 September 2023

Keywords:

Machine Learning,
Prediction,,
Regression,
Educational Success

Doi: 10.24012/dumf.1322273

ABSTRACT

Today, machine learning methods are used effectively and show high performances and effective results in many areas. For this reason, the application of machine learning algorithms has become more widespread in various industries in recent years. With the inferences that can be obtained from machine learning models, many problems can be predicted and solved. The aim of the study is to use the data obtained from the questionnaire made with secondary school students; To introduce a machine learning model that will predict educational success and to determine the factors that may negatively affect the student. The questions of the questionnaire were created by investigating in literature that the factors that may affect the success of the student. Within the scope of the study, data were collected from 519 different students studying in various secondary schools through a questionnaire consisting of 13 questions within the scope of the law on the protection of personal data. This data has not been shared with any institution and confidentiality is preserved. For modeling after dataset preprocessing and visualization, K-Nearest Neighbor (K-NN) , Random Forests (RO) , Linear Regression, Bagged Trees Regression (BTR), Gradient Boosting Regressor (GBM - Gradient Increasing Regression)) and Decision Trees (CA) algorithms were used. In the study, after the data preprocessing steps were carried out, the model created by the use of machine learning methods was used to predict the educational success of the student over the Turkish grade. In the study, the Turkish lesson was chosen as the dependent variable because of the determination of the lesson selection, the fact that the mother tongue is Turkish and Turkish lesson is taken every semester since the education life. As a result of the study, the random forest method was the most effective method with an accuracy rate of 0.88 and an R-Square value of 0.98. The most important factors affecting the educational status of the student were found to be the correlation between them depending on the Turkish grade, family income and study hours.

* Sorumlu Yazar

Giriş

Türkiye'de eğitim sistemi 12 yıllık zorunlu eğitim olarak 3 kademeye ayrılmıştır. Şu anda 4+4+4 eğitim sistemini kullanılmaktadır. Son yıllarda eğitim sistemi birçok kez değiştirilmiştir. Eğitim sistemindeki değişiklikler bazı konularda öğrencileri olumlu ve/veya olumsuz yönde etkilemiştir.

Lise eğitimine geçmeden önce son basamak 8.sınıftır. Öğrenci bu aşamada hayatını etkileyecek bir döneme girmekte olup, lise eğitimi sayesinde üniversiteye girişe hazır hale gelecektir. Öğrenci üzerinde önceki senelere göre daha fazla baskı olabilmektedir. Sosyal çevre, aile ilişkisi, öğretmeniyle etkileşimi, özel ders desteği, ailelerin geliri, oyun oynama ve televizyon izleme süresi, uyku düzeni ve bunun gibi birçok etken öğrencinin başarı durumunu doğrudan etkileyebilmektedir [1]. Çelenk'in bulguları, ailelerinden eğitim desteği alan ve okulla yakın bağları olan evlerdeki çocukların okumada daha iyi performans gösterdiği yönündedir. Ailenin öğrenciye karşı ilgisizliği, öğretmen tarafından verilen ödevlerde öğrencinin yetersiz kaldığı noktalarda gerekli yardımlaşmanın sağlanmaması da başarıyı etkilemektedir. Öğretmenler her ne kadar velilerin daha özverili olmalarını isteseler de gerek ailenin eğitim konusundaki yetersizliği gerek iş yoğunluğu kaynaklı sebeplerden öğrenci bu konuda eksik kalmaktadır [2]. Çelenk, aileleri eğitime destek veren, okulla düzenli iletişim halinde olan ve ortak programları kabul eden çocukların daha başarılı olduğunu araştırmasında ortaya koymuştur. Çalışma ortamı, maddi yetersizlik, ailenin kalabalık olması öğrencinin başarısızlığını ortaya çıkaran etkenlerden olabilmektedir.

Literatürde Makine Öğrenmesi yöntemleri araştırılarak benzer algoritmalarla yapılan çalışmalar incelenmiştir. Yavuz, çalışmasında, K-NN, Rastgele Orman, Doğrusal Regresyon modellerini kullanarak enerji tüketimi üzerinde tatminkâr sonuçlar üretmeyi ve enerji yönetiminde kolaylık sağlamayı amaçlamıştır [3]. Elde edilen sonuçlar incelendiğinde 0,0067 MSE değeri ile KNN yönteminin en başarılı algoritma olduğunu ifade edilmiştir. Gök, makalesinde, öğrencilerin yaşam koşullarının ve sosyal çevrelerinin Türkçe notunu nasıl etkilediğini belirlemek için ilk olarak 6, 7 ve 8. sınıflardaki öğrencilere 24 maddelik demografik özellikler içeren bir anket uygulanmış ve etkisini değerlendirilmiştir. Çalışmanın sonunda Genel Başarı ortalamasına ait puan tahmininde ortalama karesel hatanın karekökü hata metriği ile değerlendirmiştir. Bu metrik baz alındığında en iyi sonucu veren algoritma 10.68 hata oranına göre Rastgele Orman algoritması olmuştur [4]. Akşehir ve ark. çalışmasında denetimli öğrenme yöntemlerinden Çoklu Lineer Regresyon, Rastgele Orman ve Karar Ağaçları kullanmıştır. Sonuç olarak, borsa endeks tahminlerinden yola çıkılarak, banka hisse senetlerinin bir gün sonraki kapanış değeri tahmininde oldukça başarılı olduğunu, yapılan analizlerden çıkan R – kare değeri göz önünde bulundurularak kullanılan tüm metotlardan %98 oranında başarı elde ettiğini ortaya çıkarmıştır [5]. Sevli, çalışmasında göğüs kanseri teşhisinde 5 farklı makine öğrenmesi yöntemi kullanarak algoritmaların performans karşılaştırmasını yapmıştır. Çalışmasında Destek Vektör Makinesi (DVM), Naive Bayes (NB), Rastgele Orman, K –En Yakın Komşu ve Lojistik Regresyon (LR) algoritmalarının test başarılarını karşılaştırmıştır. Göğüs kanserinin teşhisinde, %98.24 doğruluk oranı ile en iyi performansı gösteren Lojistik Regresyon algoritması olduğunu belirlemiştir [6]. Yağcı, çalışmasında, lisans öğrencilerinin ara sınav notları kaynak olarak göstermiştir. Rastgele Orman, KNN, Destek Vektör Makineleri, Lojistik Regresyon, Naive Bayes algoritmaları kullanılmıştır. Tahminler 3 tip parametre (ara sınav notları, bölüm verileri, fakülte verileri) kullanılarak yapılmıştır. Algoritmaların sonuçları birbirine yakın çıkarak %70-%75'lik bir doğruluk oranı göstermişlerdir**.** Random Forest algoritması, %75.2 başarı oranı ile en iyi performansı gösteren yöntem olarak belirlenmiştir [7]. Cruz-Jesus ve ark. çalışmasında, yaş, cinsiyet, derse katılım, internet erişimi, bilgisayara sahip olma ve ders sayısı gibi 16 öznitelik ve 110627 gözlem ile öğrenci akademik performansını tahmin etmek istemiştir. Rastgele Orman, KNN, Lojistik Regresyon, Destek Vektör Makineleri algoritmalarını kullanarak öğrencilerin performansını %50 ile %81 arasında değişen doğruluk oranlarıyla tahmin etmişlerdir. Logistik Regresyon %51.2 doğruluk oranı ile en düşük performansı gösteren algoritma iken Destek Vektör Makineleri %81.1 ile en yüksek performansı gösteren algoritma olmuştur [8]. Fernandes ve ark. Sürdürdükleri çalışmalarında, sınıf kullanım ortamı, cinsiyet, yaş, öğrenci yardımı, şehir, mahalle, not, devamsızlık gibi özellikleri barındıran demografik bir model geliştirmişlerdir. Bu etkenlere bağlı olarak öğrencinin akademik başarısını tahmin etmeyi amaçlamışlardır. Gradient Boosting Machine algoritmasını kullanarak 2 veri seti üzerinde çalışmışlardır. Birinci veri seti 19000, ikinci veri setinde 19834 gözlem sayısı bulundurarak modelleme yapmışlardır. Çalışmalarının sonunda algoritmanın, birinci veri setindeki başarısı %85.9, ikinci veri setinde ise %91.9 başarı oranına sahip olduğunu gözlemlemişlerdir [9]. Xu ve ark. çalışmasında, üniversite öğrencilerinin internet kullanım davranışları ile akademik performansları ve makine öğrenimi yöntemleriyle öğrencilerin performansını tahmin etmeyi amaçlamışlardır. Öğrencilerin internet kullanım süresini, internet bağlantı sıklığını, internet trafik hacmini ve çevrim içi zamanı değerlendirmiştir. Karar Ağaçları ve Destek Vektör Makineleri algoritmalarını kullanarak, 4000 gözlem sayısı elde etmişlerdir. Destek Vektör Makineleri yöntemi %73 başarı performansı ile en tatminkâr algoritma seçilmiştir [10]. Hofait ve ark. çalışmasında, cinsiyet, Uyruk, Eğitim, Önceki eğitim, matematik, burs, başarı gibi demografik özellikleri öznitelik olarak belirlemiştir. Başarısız öğrencileri tespit etmeyi amaçlamışlardır. Rastgele Orman, Yapay Sinir Ağları, Logistik Regresyon algoritmalarını kullanarak 2244 gözleme sahip bir veri setiyle çalışmalarını sürdürmüşlerdir. Yapay Sinir Ağları yöntemi %70.4 ile en düşük başarı performansını gösterirken, Rastgele Orman %90 başarı oranı ile en yüksek performansı göstermiştir [11]. Chui ve ark. çalışmalarında, Destek Vektör Makineleri algoritmasını kullanarak risk altındaki öğrencilerin akademik başarı performanslarını tahmin etmeyi amaçlamışlardır. Çalışmanın sonunda, 32.593 gözlem sayısına sahip veri seti ile model performansını %93.5 olarak belirlemişlerdir [12]. Nieto ve ark. makalelerinde,

ortaokul öğrencilerinin performansını tahmin etmek için Destek Vektör Makineleri ve Yapay Sinir ağları algoritmalarının performanslarını karşılaştırmıştır. Veri setinde Microsoft Showcase School tarafından toplanan 5520 öğrencinin performans verileri bulunmaktadır. Destek Vektör Makinesinin %84.54 oranında başarı performansı olduğunu ortaya çıkarmışlardır [13]. Ahmad ve ark. öğrencilerin akademik başarılarının risk altında olup olmadığı sorusuna cevap aramıştır. MPNN yöntemini seçerek, 300 gözlem sayısına sahip veri setiyle çalışmıştır. Demografik olarak önceki derece notları, ev ortamı, çalışma alışkanlıkları, öğrenme beceri özelliklerini değerlendirmiştir. Çalışmanın sonunda, %95 model performansı elde etmişlerdir[14]. Bu çalışmada öğrencinin eğitim başarısını etkileyecek etmenler üzerine yoğunlaşılarak anket sorularının oluşturulmasında araştırma makaleleri ve bilimsel çalışmalar incelenmiştir. Dam, araştırmasında, otobiyografi tekniği ile öğrencilerden veri toplamıştır. Öğrencilerin aileleriyle yaşadıkları iletişimsizlik, diğer kişilerle kıyaslanma durumu, ailelerinin fazla beklenti içerisine girdikleri, aile içi huzursuzluk, boşanma, ailede yaşanan ölüm, uygun çalışma ortamının bulunmaması, ekonomik durumlar başarıya etki eden faktörlerden olduğu sonucuna ulaşmıştır [15]. Aslanargün ve ark. çalışmasında, öğrencilerin başarısını etkileyen aile eğitimi, öğrenci cinsiyeti, ekonomik durum, kardeş sayısı gibi faktörlerin başarıyla nasıl bir ilişki içerisinde olduğu yönünde bir araştırma yapmıştır. Öğrencilere ulaştırılan anket soruları ile veriler toplanmış ve analizi gerçekleştirilmiştir. Araştırmanın sonunda, tek çocuk veya çok kardeşe sahip olma durumunun, 2 kardeşe sahip olma durumuna oranla daha az başarılı olduğu, gelirin öğrenci başarısında pozitif etkisinin olduğu, cinsiyet faktörünün erkekler için olumlu yönde olduğu sonuçlarına ulaşmıştır [16].

Literatürde yer alan benzer araştırmalardan farklı olarak bu çalışmada, plot ders olarak seçilen Türkçe dersi için öğrencinin başarım durumunu doğrudan olumsuz etki eden etmenlerin tespit edilmesidir. Bunun yanında, makine öğrenmesi yöntemleri ile modellenen sisteme sunulan yeni öğrenci bilgileri ile öğrencinin Türkçe notunun tahmin edebilmesi amaçlanmıştır. Ayrıca eğitim başarımı öngörüsü ve başarımı olumsuz etkileyen unsurların tespiti konularında makine öğrenmesi yöntemlerinin etkin olarak kullanımının ortaya konması çalışmadaki hedeflerdendir.

Çalışmayı kapsayan 13 farklı anket sorusu bulunmaktadır. Anket soruları ve öğrencinin başarısını etkileyen parametreler literatürde yer alan araştırmalar incelenerek belirlenmiştir. Genel olarak ankette; aile durumu ve geliri, özel ders durumu, ders çalışma, oyun oynama, uyku düzeni, kardeş olma durumu, kendine ait oda durumu gibi sorulara yer verilmiştir. Ortaokul 8.sınıf öğrencileri üzerinden yapılan anket ile 519 gözlem elde edilmiştir. Buna bağlı olarak oluşturulan veri setinde 13 sütun ve 520 satır bulunmaktadır.

Çalışmada veriyi modellemek için benzer problemlerde en sık kullanılan makine öğrenmesi yöntemleri olan Çoklu Lineer Regresyon, Rastgele Orman, K-En Yakın Komşu ve Torbalanmış Ağaçlar, Karar Ağaçları ve Gradyen Arttırıcı Regresyon yöntemleri kullanılmıştır. Makine öğrenmesi algoritmaları ile modellenen sistem, Türkçe notunu tahmin etmesi beklendiği için 13 değişken içerisinde Türkçe notu hedef değişken ve model çıktısı olarak seçilmiştir. Verisetinin %80'i eğitim seti, %20'si test seti olarak kullanılmıştır. Eğitim sonrasında modelin tahmin başarısını değerlendirmek için model doğrulama (tuning) işlemi yapılmıştır. Bu çalışmada kullanılan makine öğrenmesi yöntemleri açıklanmış, veri setinde yapılan ön işleme adımları sonrasında elde edilen temiz verilerle çalışmada kullanılan 6 makine öğrenmesi algoritması ile modellenmiştir. Son olarak; öğrencinin cevapladığı anket sorularına karşılık sistem Türkçe notunu tahmin edebilecek bir modelleme ortaya çıkarmıştır.

Materyal ve Metot

Veri seti

Çalışmada kullanılan veri seti ortaokul öğrencilerine uygulanan 13 soruluk anket aracılığı ile oluşturulmuştur. Jupiter Notebook'ta modelleme ve manipülasyon işlemlerinin yapılabilmesi için .csv formatıyla işlenmiştir. Anket sorularını cevaplayan 519 öğrenciden 273 tanesi kız, 246 tanesi erkek öğrenci olarak belirlenmiştir. Veri setine ait öznitelik bilgileri Tablo1'de gösterilmiştir.

Çalışmada Çoklu Lineer Regresyon, Rastgele Orman, K-NN ve Torbalanmış Ağaçlar, Karar Ağaçları ve Gradyen Arttırıcı Regresyon yöntemleri ile modellenmiştir. Veri setindeki gözlem sayısı göz önünde bulundurulduğunda modelin daha iyi öğrenebilmesi ve test verilerinin daha güvenilir sonuçlar verebilmesi için %80 eğitim, %20 test seti olarak veri kümesi ayrılmıştır. Modellerin tahmin başarısı RMSE, MSE, R-kare metrikleriyle değerlendirilmiştir. Model performanslarını daha doğru ve objektif değerlendirebilmek için K-Cross Validation (Çapraz Doğrulama) yöntemi kullanılmıştır.

Tablo 1.Özniteliklerin Anket Sorularına Karşılıkları

Anket Soruları	Öznitelik Bilgileri İngilizce Karşılıkları	Öznitelik Bilgileri Türkçe Karşılıkları
Cinsiyetiniz	Sex	Cinsiyet
Kardeşiniz var mı ?	Sibling	KardeşSayısı
Özel eğitim alıyor musunuz?	Coaching	ÖzelDers
Kendine ait bir odanız var mı?	HaveRoom	KendineAitOda
Anneniz ve babanız ayrı mı?	FamSep	AileAyrılıkDurumu
Anneniz veya babanız hayatta mı?	FamilyAlive	AileYaşamaDurumu
Üvey anneniz veya babanız var mı?	StepFM	ÜveyAnneBaba
Hafta sonları ders çalışıyor musunuz?	WStudying	HaftasonuDersÇalışma
Günde kaç saat ders çalışıyorsunuz?	Studying	DersÇalışmaSüresi
Günde kaç saat ders uyuyorsunuz?	Sleeping	Uyku
Ortalama kaç saat oyun oynuyorsunuz?	Playing	Oyun
Aile geliriniz nedir?	Income	Gelir
Türkçe Notunuz nedir ?	TurkNote	TürkçeDersNotu

Makine Öğrenmesi Ve Yapay Zeka

Makine öğrenimi, 1959'da yapay zekada sayısal öğrenme ve model tanıma çalışmalarından doğan, esasen bilgisayar biliminin bir alt alanıdır. Makine öğrenimi, bir yapı işlevi olarak verilere dayalı tahminler yapabilen algoritmalar oluşturma sistemidir. Bu tür algoritmalar statik program yönergelerini sıkı sıkıya takip etmek yerine, veriye dayalı tahminler ve kararlar almak için örnek girdilerden modeller oluşturarak çalışır [17]. Makine öğrenimi temel bir sorundan doğmuştur. Bilgisayarların bizim çalışma şeklimizin ötesine geçip belirli işlemleri kendi başlarına gerçekleştirmeyi öğrenmeleri mümkün müdür? Geliştiricilerin verilerinden öğrenen kuralları manuel olarak oluşturması yerine bilgisayar verilerine bakar [18].

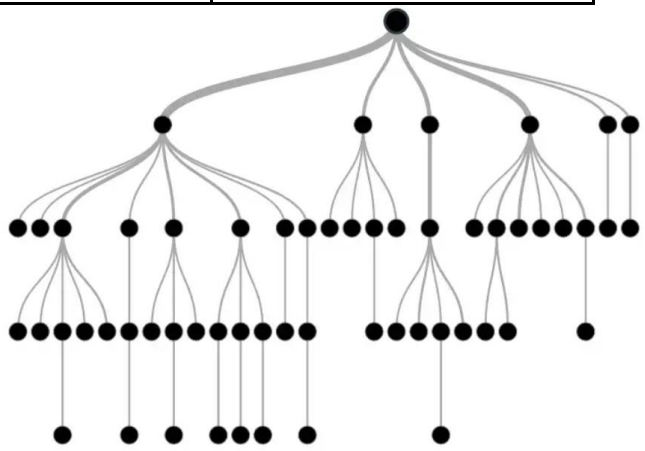
Makine Öğrenmesi yaklaşımları; Denetimli Öğrenme, Denetimsiz Öğrenme, Pekişmeli Öğrenme, Yarı Gözetimli Öğrenme olarak ayrılmıştır.

Makine Öğrenmesi Algoritmaları

Makine öğrenmesinde kullanılan birçok algoritma mevcuttur. Bu çalışmada sonuç değişkeni (Türkçe Notu) sürekli değişken (sayısal) olduğu için makine öğrenmesi algoritmalarından regresyon modelleri seçilmiştir.

Karar Ağaçları

Sınıflandırma ve regresyon yöntemlerinin bir metodu olan karar ağacı, yorumlama kolaylığı olması, veri tabanı sistemleriyle entegrasyon kolaylığı ve yüksek güvenirliği nedeniyle sınıflandırma yöntemleri arasında en yaygın kullanılan yöntemdir [19]. Karar ağaçları algoritmasının kök, düğüm ve yapraklarının şematik gösterimi Şekil 1'de gösterilmiştir.



Şekil 1. Karar Ağaçları Şematik Gösterimi

K-En Yakın Komşu

K-en yakın komşuluk algoritması, uygulaması kolay gözetimli öğrenme algoritmalarındandır [20]. Yöntem uygulanırken öncelikle en yakın sınıflı veri sayısını ifade eden bir K değeri belirlenmektedir. Sınıf değeri bilinmeyen yeni bir değer tahmin için sisteme sunulduğunda, K kadar en yakın sınıf değeri bilinen elemanın bu değer arasındaki uzaklık hesabı ile sınıf tahmini yapılmaktadır. Uzaklık hesaplama işlemlerinde genelde Öklid uzaklık hesaplaması kullanılır.

Euclidean (Öklidyen) uzaklık hesaplamasında kullanılan k değeri boyut sayısını, x_i veri seti içerisindeki değerleri, y_i tahmin edilecek değerleri temsil eder.

$$Öglid = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (1)$$

Çoklu Lineer Regresyon

İki veya daha fazla bağımsız değişken (X) ile bir bağımlı değişken (y) arasındaki ilişkiyi tahmin etmek için kullanılan yöntemdir.
Denklemde y değişkeni bağımlı değişkenin tahmin edilen değerini, β_0 y kesişim sabitini , $\beta_1 X_1$ birinci bağımsız değişkeni , $\beta_n X_n$ son bağımsız değişkenin regresyon katsayısını , ϵ model hatasını temsil etmektedir.

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \qquad (2)$$

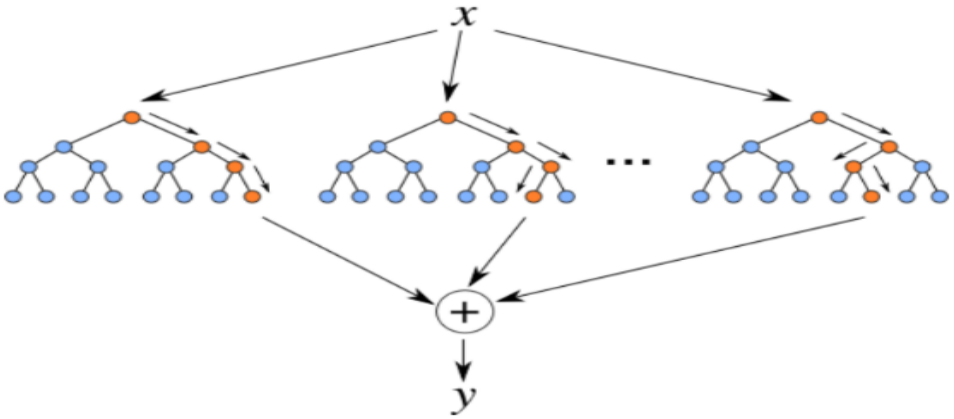
Çoklu doğrusal regresyon modelinde, bazı bağımsız değişkenlerin birbiriyle ilişkili içinde olması mümkündür. Bu nedenle regresyon yönteminde geliştirmeden önce bunları kontrol etmek gerekmektedir. Eğer iki bağımsız değişken çok yüksek oranda ilişkiliyse o zaman regresyon yönteminde sadece biri kullanılmalıdır.

Gradyen Arttırıcı Regresyon- Gradient Boosting Machines

Makine öğrenmesine ait bu yöntem hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılan bir algoritmadır. Bu yöntemin yaklaşımı, tek bir karar ağacının yeteri kadar kuvvetli olmadığı yönündedir. Bir karar ağacı yöntemi uygulanır. Ortaya çıkan hata ile yeni bir karar ağacı oluşturulur. İşlem hata minimuma ininceye kadar devam ettirilir.

Rastgele Orman

Rastgele Orman, sınıflandırma, eğitim aşamasında çok sayıda karar ağacı üreten ve sorunun türüne bağlı olarak sınıfı veya sayıyı tahmin eden, sınıflandırma, regresyon, ve diğer görevler için eş zamanlı öğrenme yöntemidir.[21] Rastgele Orman yöntemi birden fazla ağacın bir araya gelerek en yüksek puan alan yöntemin seçilmesidir ve şematik olarak Şekil'3 te gösterilmiştir.



Şekil 2. Rastgele Orman Algoritmasının Yapı

Torbalanmış Ağaçlar-Bagged Trees

Birkaç karar ağacından gelen tahminleri birleştiren makine öğrenmesi algoritmasıdır. Bir karar ağacının en büyük sorunu yüksek varyansa sahip olmasıdır. Yani veri setindeki herhangi küçük bir değişik modelde veya yapılacak tahminlerde büyük
değişikliklere neden olabilir. Varyansı en aza indirmek için Torbalanmış Ağaçlar- Bagged Trees algoritmaları kullanılabilir.

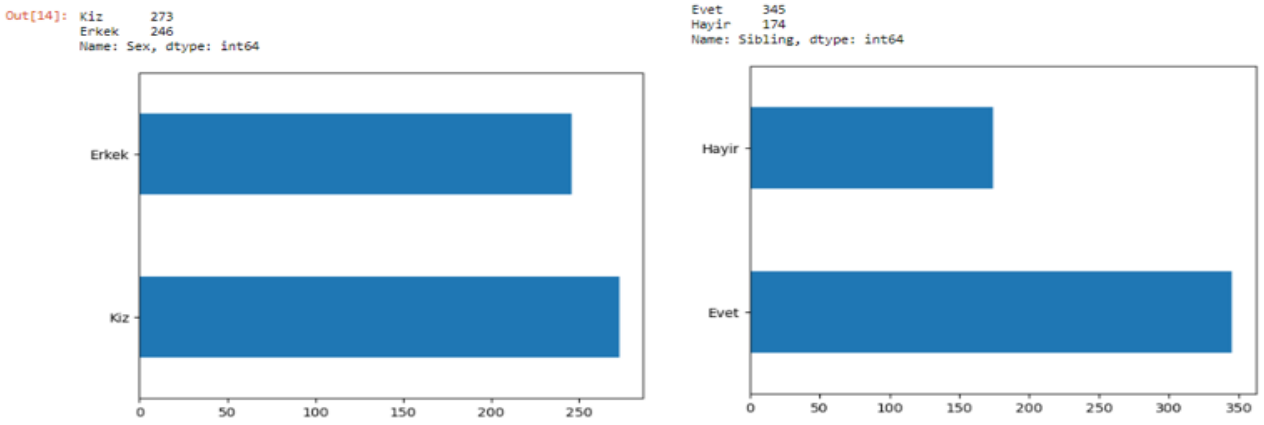
Deneysel Çalışmalar

Anket sorularının her biri için değişken ismi tanımlanmıştır. Veri seti 8 kategorik ve 5 sayısal değişken içermektedir. Veri seti içerisinde performansı etkileyecek durumlar göz önünde bulundurulmuş; gerekli ön işlemle adımları uygulanmıştır. Eksik bir değerin olması modelin başarısını etkileyen en kritik etkenlerden biridir. Veri seti içerisinde eksik bir veri olup olmadığı Python programlama dili ile, isnull parametresi ile kontrol edilmiş ve eksik veriye rastlanmamıştır.

Şekil 3'te "value_counts().plot.barh()" parametresi ile Veri setinden örnek olarak cinsiyet (Sex) ve kardeşe sahip olma (Sibling) kategorik değişkenleri grafiksel olarak betimlenmiştir. Bu grafiklere göre anket sorularını cevaplayan 519 öğrenciden 273 tanesi kız öğrenci, 246 tanesi erkek öğrencidir. Ayrıca, 343 öğrenci kardeşe sahip, 174 öğrencinin ise kardeşi olmadığı sonucuna varılmıştır. Böylelikle anketi cevaplayan daha fazla kız öğrenci olduğu ve kardeşe sahip olma öğrencilerin çoğunlukta olduğu gözlemlenmiştir.

Veri seti üzerinde aykırı gözlem analizi yapılmıştır. Grubbs'a göre aykırı gözlem; "Aynı örneklem içindeki diğer gözlemlerden belirgin derecede farklı olan veya sapma gösterendir". Bu sapmalar ve farklılıklar modelin elde edeceği başarıyı etkilemektedir [22]. Bu nedenle çalışmada, kutu grafiği (boxplot)yöntemi ile aykırı gözlem analizi yapılmıştır. Kutu grafiği yöntemi; veri çeyreklerini

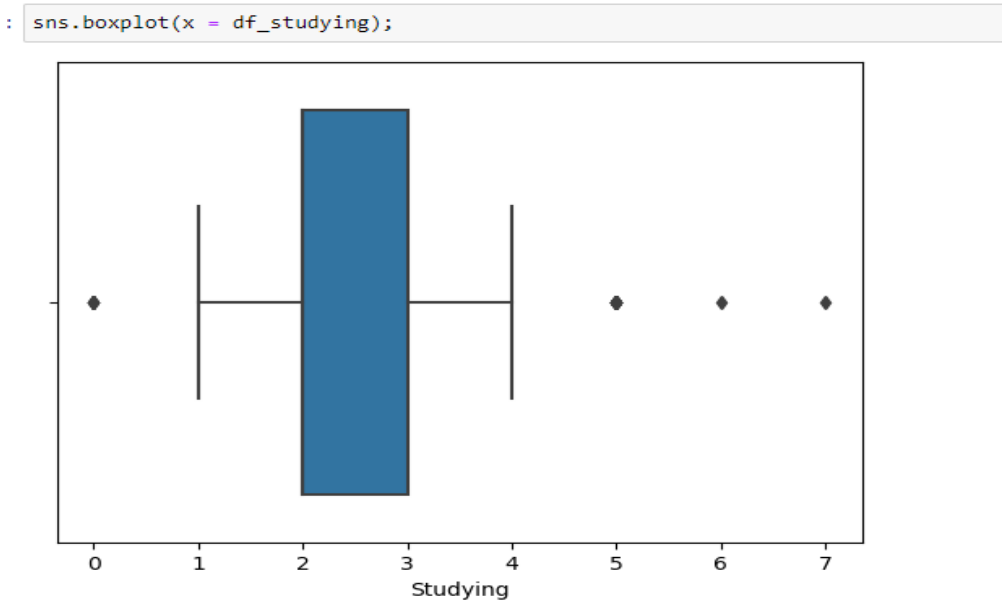
(yüzdelikleri) ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğin görsel olarak dağılımını



göstermek için kullanılmıştır.

Şekil 3. Cinsiyet(Sex) Ve Kardeşe Sahip Olma(Sibling) Değişkenlerinin plot.barh() İle Gösterimi

Kutu grafiği yöntemi ile görselleştirilen değişkenler arasında studying(Ders Çalışma Süresi) ve income(gelir) değişkenlerinde aykırı gözleme rastlanmıştır. Baskılama yöntemi ile alt sınır değerleri belirlenmiş ve alt sınır değerinden aşağıda kalan değerler baskılanmıştır. Bununla birlikte modelin başarısında 0.01 değerinde artış gözlemlenmiştir. Şekil 4'te kutu grafiği uygulanması gereken aykırı değere sahip studying (Ders Çalışma Süresi) değişkeni gösterilmiştir. Veri setindeki değişkenlerin sahip olduğu ilk 10 değer örnek olarak Tablo 2'de gösterilmiştir.



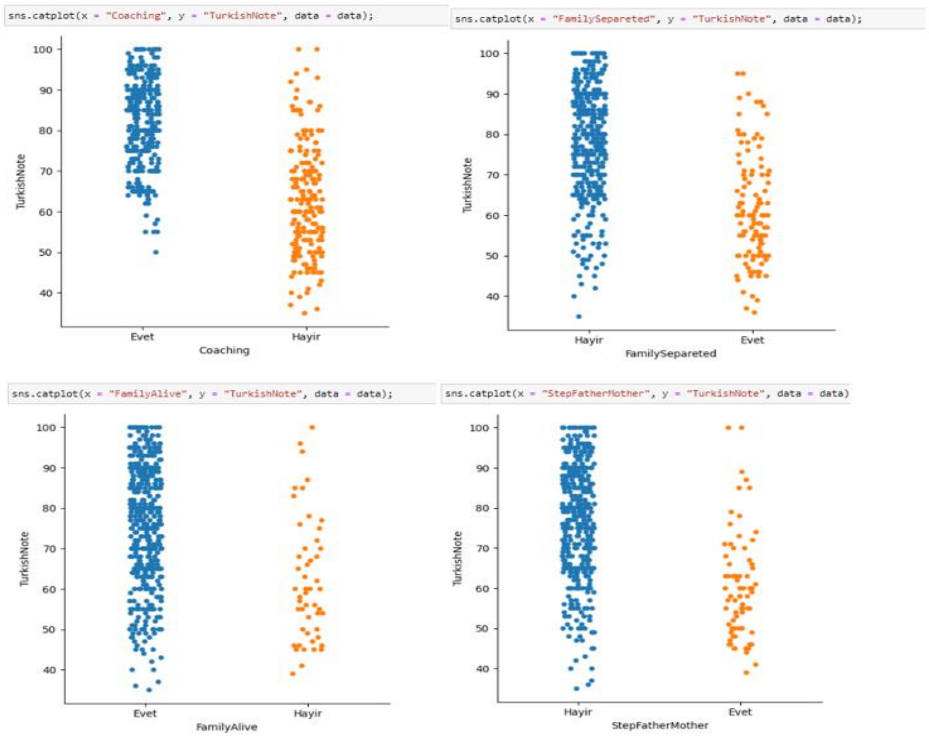
Şekil 4. Aykırı değere sahip Ders Çalışma Süresi (studying) değişkeninin Boxplot yöntemiyle gösterimi

Tablo 2. Veri Setinin Detayları

Sex	K	K	K	Er	Er	K	Er	K	K	Er
Sibling	E	E	H	E	E	H	E	H	E	E
Coaching	E	H	H	E	H	E	H	H	H	H
HaveRoom	E	H	E	H	E	E	E	E	H	H
FamSep	H	H	H	E	H	E	H	E	H	H
FamilyAlive	E	E	E	E	E	E	E	E	E	E
StepFM	H	H	H	E	H	H	H	E	H	H
WStudying	E	H	E	H	H	E	E	E	E	H
Studying	2	1	3	4	3	3	5	2	4	2
Sleeping	9	10	8	9	6	9	7	8	8	7
Playing	1	1	0	3	5	0	1	1	3	4
Income	9000	8900	11000	9200	9800	8500	11000	7000	10000	6000
TurkNote	85	67	94	50	72	87	80	76	85	56

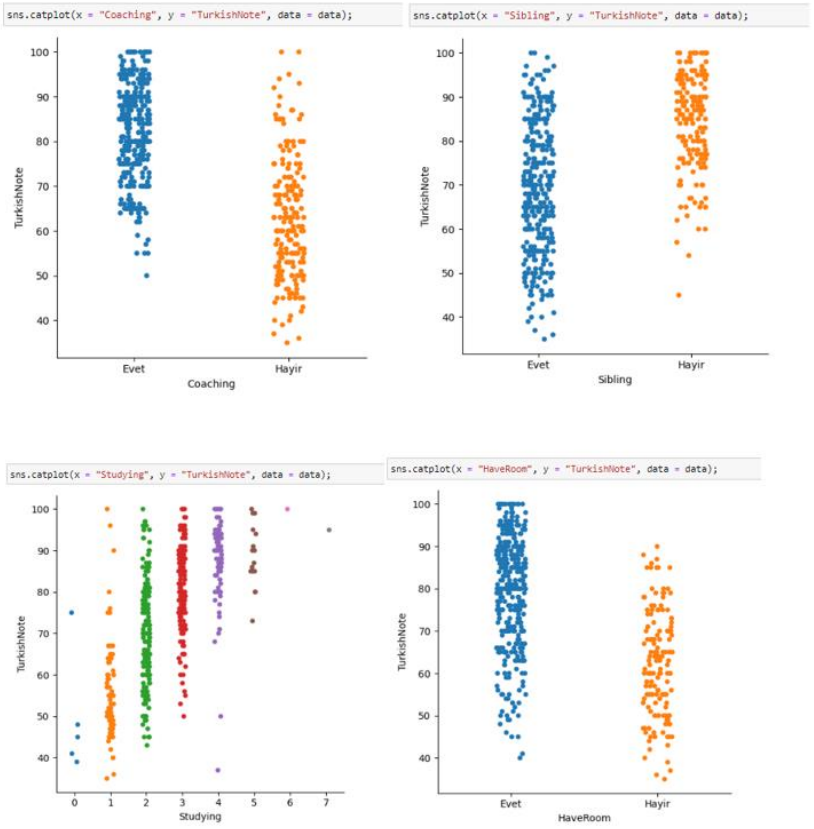
Veri setinin içeriğini daha iyi analiz etmek için modelleme aşamasına geçmeden önce değişkenler arasında ilişki grafikleri kurulmuştur.Şekil 5'te Seaborn kütüphanesine ait catplot parametresi kullanılarak özel eğitim alma durumu (Coaching), ailenin ayrılık durumu (FamilySeperated), ailenin hayatta olma durumu (FamilyAlive) ve üvey anne babaya sahip durumu (StepMotherFather) değişkenleri "sns.catplot()" parametresiyle görsel olarak Türkçe notuyla ilişkilendirilmiştir. Görseller göz önünde bulundurulduğunda;

i)Veri seti içerisinde ağırlıklı olarak özel eğitim desteği alan öğrencilerin notlarının 70 ile 100 aralığında olduğu ,
ii)Ailesi ayrı olmayan öğrencilerin Türkçe notunun 60 ile 98 arasında olduğu,
iii)Ailesi hayatta olan öğrencilerin notlarının 50 ile 100 arasında yoğunlukta olduğu,
iv)Üvey anne babaya sahip olmayan öğrencilerin 60 ile 100 arasında not aldığı,
Sonuçlarına ulaşılmıştır.



Şekil 5. Catplot parametresi ile kategorik değişkenlerin gösterilmesi

Şekil 6'da Odaya sahip olma (HaveRoom), özel eğitim alma (Coaching), kardeşe sahip olma (Sibling), ders çalışma (Studying) değişkenlerinin catplot parametresi ile görsel olarak betimlenmesi gösterilmiştir. Görsel yorumlandığında;
i)Özel eğitim alan öğrencilerin not aralığının 65 ile 100, almayan öğrencilerin 45 ile 75 arasında olduğu,
ii)Kardeşe sahip öğrencilerin ağırlıklı olarak 45 ile 90, olmayan öğrencilerin 75 ile 100 arasında ağırlıklı olduğu,
iii)Genellikle 2,3,4 saat ders çalışan öğrencilerin notlarının 55 ile 100 arasında olduğu,
iv)kendi odasına sahip öğrencilerin notlarının 65 ile 100, olmayan öğrencilerin 50 ile 75 arasında olduğu sonucuna varılmıştır.
Türkçe Dersi notu bağımlı değişkeniyle sayısal değişkenlerin arasındaki korelasyon hesaplanmış;negatif yönde etki eden 2 değişken Oyun, hafta sonu ders çalışma (Playing, WeekendStudying) veri setinden çıkartılmıştır. Veri seti içerisindeki kategorik değişkenleri sayısal değişkenlere dönüştürebilmek için Label Encoder olarak adlandırılan Etiket Kodlayıcı yöntemi kullanılmıştır. Evet-Hayır ve Erkek-Kız kategorik değişkenleri bu yöntemle 0-1 olarak dönüşüme uğramıştır.
Çalışmada, makine öğrenmesi yöntemleri destekli modelleme için uygulanan işlem adımları Şekil 7' de gösterilmiştir.



Şekil 6.Catplot parametresi ile kategorik değişkenlerin gösterilmesi

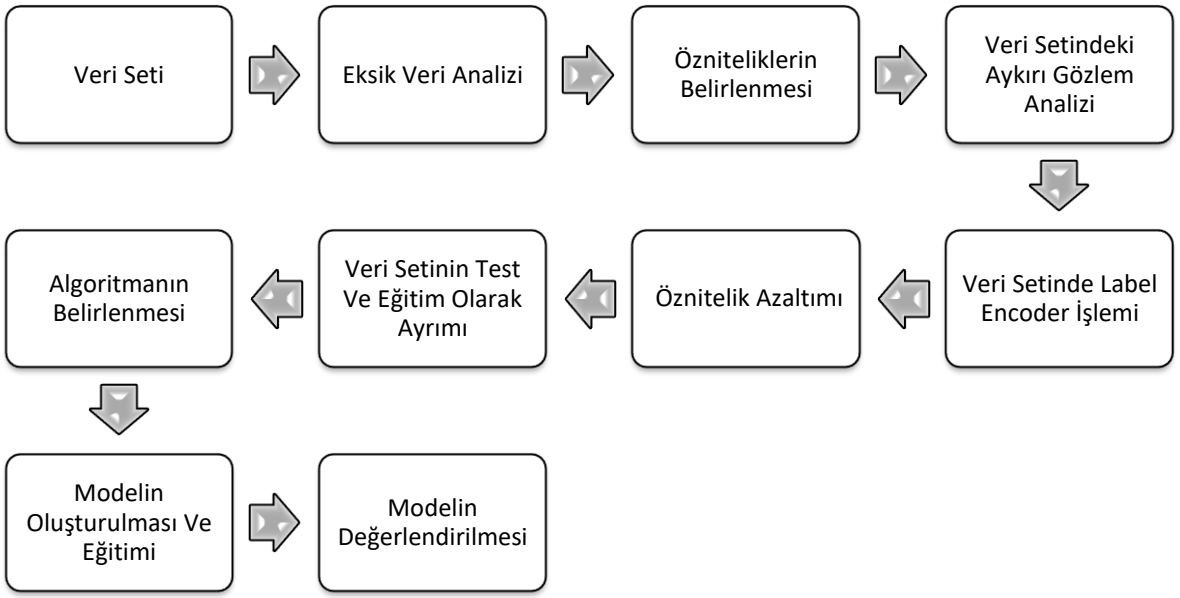
Çalışmada 6 adet denetimli öğrenme türü olan Rastgele Orman, Çoklu Lineer Regresyon, Karar Ağaçları, Torbalanmış Ağaçlar, Gradyen Arttırıcı Regresyon, K-NN algoritmaları kullanılmıştır. Bu algoritmalarının seçilmesinin nedeni, yapılan literatür taraması sonucunda en çok tercih edilen ve model başarı performansları yüksek olacağı öngörülen yöntemler olmasından kaynaklıdır. Kullanılan algoritmalar model başarı yüzdeliği, başarı metrikleri RMSE, MSE, R^2 'ye göre değerlendirilmiş ve karşılaştırılmıştır.

RMSE değeri, bir veri kümesindeki öngörülen $\hat{y}_i$ değerleri ile gerçek y_i değerleri arasındaki ortalama kare farkının karekökünü hesaplamaktadır. RMSE düşükse, model veri kümesine o kadar iyi uymaktadır. RMSE değeri , MSE değerinin karekökü'dür.

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

R^2 değeri, verilerin regresyon hattına ne kadar yakın olduğunun istatistiksel bir metriğidir. Belirleme katsayısı olarak da bilinmektedir. Özetlemek gerekirse R-kare, doğrusal regresyon metodları için uygunluk ölçüsüdür. Explained Variation değeri toplam hatalar karesini, Total Variation değeri tüm toplam kareleri temsil etmektedir.

$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation} \quad (4)$$



Şekil 7. Makine Öğrenmesi Yöntemlerinin Uygulanması Aşamaları

Bulgular

Denetimli öğrenme algoritmalarının her biri için bir model oluşturulmuş ve eğitilmiştir. 6 adet algoritmanın; model başarıları, R-kare değerleri (model uyumluluğu), Çapraz Doğrulama değerleri, RMSE ve MSE metrikleri Tablo 3'te karşılaştırılmıştır.

Tablo 3 göz önünde bulundurulduğunda en yüksek model performansını gösteren model Rastgele Orman yöntemi olmuştur. Algoritmaların RMSE ve MSE değerleri de sisteme hesaplatılmıştır. Tablo 3'e göre RMSE değeri düşük olan algoritmanın performansı yüksektir. RMSE metrik değeri göz önünde bulundurulduğunda Rastgele Orman algoritması en iyi performansı gösteren model, K-NN algoritması en düşük performansı gösteren model olarak belirlenmiştir. Regresyon hattına yakınlığı temsil eden R-Kare (R2) değerlerinin karşılaştırılması tablo 3'te gösterilmiştir.
Tablo 3'te model uyumluluklarına bakıldığında Rastgele Orman algoritması regresyon hattına en yakın yöntemdir. K-NN algoritması ise 0.29 R-kare değeri ile regresyon hattına en uzak yöntem olmuştur. Modelin performansını değerlendirmek ve genelleştirme yeteneğini test etmek için model doğrulama yöntemi olarak çapraz doğrulama Çapraz Doğrulama veri setini daha küçük alt katmanlar halinde bölmektedir. Bu alt katmanlar dönüşümlü olarak kullanılır.

Karar Ağaçları ve GBM algoritmalarının model başarısı yüksektir. Ancak R-kare değerinin düşük olmasından kaynaklı yeni girdiler sisteme verildiğinde yanıltıcı çıktı üretebileceğinden bu algoritmalar tercih edilmemiştir.

K-NN algoritması düşük başarı ve regresyon hattına sahiptir. İyi sonuç çıktısı alınması oldukça güçtür. Algoritma bu sebeple tercih edilmemiştir. BTR ve ÇLR algoritmalarının başarı yüzdesine bakıldığında yüksek olsa da, model uyumluluğu en yüksek algoritma değildir. Daha iyi tahmin üretebilecek yeni girdilerle uyum sağlayabilecek bir algoritma modeli sonucuna ulaşıldığı için BTR ve ÇLR yöntemleri kullanılmamıştır.

Yapılan analizler ve değerlendirmeler sonucunda çalışmada, %88 başarı oranı, %98 model uyumluluğu ve en düşük RMSE değerine sahip olduğu için en etkin sonuç üreten algoritma rastgele orman yöntemi olmuştur.
Kullanılan diğer 5 algoritmanın Rastgele Orman modeli kadar başarılı olmamasının sebebi veri setinde kullanılan gözlem sayısının az olmasından kaynaklı olabileceği düşünülmektedir. Değişken sayısının artması modelin başarısını negatif yönde etkileyebilmektedir.

GBM yöntemi, literatürde en çok kullanılan algoritmalar kadar başarı (%77) göstermiştir. Ancak R-kare (model uyumluluk) değerinin düşük olmasından kaynaklı model seçiminde elenmiştir. Çalışmada, beklenen çıktı değeri sürekli değişken olduğundan uygulamada modele doğrusal ve doğrusal olmayan regresyon modelleri uygulanmıştır.
En iyi sonucu elde etmek için model, farklı senaryolarda defalarca test edilmiştir. Sonuç olarak Playing(Oyun) ve WeekendStudying(Haftasonu Ders Çalışma) değişkenleri modelin tahmin performansını negatif yönde etkileyen parametre olarak bulunmuş ve veri seti içeresinden çıkarılmıştır. Model bu değişkenler çıkarıldıktan sonra tekrar test edilmiştir. Test sonucunda algoritma daha yüksek performanslı sonuç üretmiştir.

Tablo 3. Algoritmaların model başarısı

	K-NN	ÇLR	RO	BTR	KA	GBM
Model Score(Model Başarısı)	0,44	0,67	0,88	0,8	0,83	0,77
RMSE	12,77	8,56	9,06	9,53	11,34	10,29
MSE	61,48	61,46	82,16	61,46	61,55	75,46
R-kare	0.29	0.98	0.66	0.61	0.34	0.33
Çapraz Doğrulama	0,49	0,7	0,92	0,83	0,86	0,79

Ayrıca literatür araştırması ile incelenen benzer çalışmalarda elde edilen doğruluk oranları kıyaslaması Tablo 4'te gösterilmektedir. Her çalışmanın farklı veri seti ve farklı parametreler ile çalışıldığı da göz önüne alındığında çalışmada elde edilen doğruluk oranları ve başarı metriklerinin literatürde geçerliliği ve uygulanabilirliği olduğu ortaya konmaktadır.

Tablo 4. Benzer çalışmaların karşılaştırılması

Çalışma	Yöntem	Doğruluk Oranı (%)
Yağcı [7]	Rastgele Orman	75.2
Jesus [8]	Destek Vektör Makineleri	81.1
Fernandes ve ark. [9]	GBM	85.9
Nieto ve ark. [13]	Destek Vektör Makineleri	84.54
Önerilen model	Rastgele Orman	88

Tartışma ve Sonuç

Bu çalışmada, öğrencilerden anket yolu ile toplanan verilerle bir veri seti oluşturulmuş; 6 farklı makine öğrenmesi algoritması kullanılarak yapılan modellemelerde en etkin sonuç üreten yöntemin Rastgele Orman algoritması olduğu ortaya konmuştur. Rastgele Orman yöntemi uygulanan model, %88 başarı oranı ve %98 R-Kare oranı elde etmiştir.

Bu çalışma Türkiye'de mevcut eğitim sistemindeki öğrencilerin hangi etkenlere bağlı olarak başarılı ya da başarısız olduğu sonucuna ulaşmak için yapılmıştır.
Elde edilen bulgular incelendiğinde, gelir durumu yüksek olan ya da daha çok ders çalışan bir öğrencinin notunun daha yüksek olabileceği sonucuna ulaşılmıştır. Veri setinde toplamda 519 gözlem vardır. Ancak veri seti büyütülürse modelin performansı pozitif yönde olacaktır. Bunun sebebi eğitim ve test için ayrılan veri setinde kullanılacak verinin çok daha fazla olmasından kaynaklıdır.

Daha büyük bir veri seti ile çalışıldığında öğrencinin notunu etkileyecek farklı etmenler belirlenebilir, ayrıca modellerin performansının yükseltilebileceği düşünülmektedir. Bu şekilde model daha iyi eğitilecek ve çok daha yüksek başarımlı tahmin sonuçları elde edilebilecektir.

Anket yolu ile elde edilen veriler 2021 aralık ayında toplanmıştır bu sebeple test aşamasında gelir değişkeni göz önünde bulundurulurken 2021 yılı baz alınarak test edilmiştir. Ortaokul öğrencilerine yöneltilen ve cevaplanması beklenen anket soruları bilimsel makaleler araştırılarak oluşturulmuştur. Çalışma ortaokul öğrencileri üzerinde yapılmıştır. Bu sebeple gelecek çalışmalar, eğitim düzeyi ortaokul kapsamından genişletilerek lise ve üniversite öğrencileri üzerinde eğitim başarısının ya da eğitim kurumlarında motivasyon etmenlerinin araştırılarak daha kapsamlı hale getirilebilmesi planlanmaktadır.

Etik Kurul Onayı ve Çıkar Çatışması Beyanı

Hazırlanan makalede veri setinin oluşturulmasında öğrencilere yöneltilen anket soruları için Beykent Üniversitesi Fen ve Mühendislik Bilimleri Bilimsel Araştırma ve Yayın Etiği Kurulu 7 üyesi,1 başkan yardımcısı, 1 başkan tarafından değerlendirilmiş ve onay alınmıştır. Hazırlanan makalede herhangi bir kişi/kurum ile çıkar çatışması bulunmamaktadır.

Teşekkür
Yazarlar, çalışmaya değerli vakitlerini ayırdıkları, çalışmaya katkıları için dergi editörlerine ve hakemlerine teşekkür etmektedir. Bu çalışma, Doç.Dr. Atınç Yılmaz'ın danışmanlığında yürütülen Deniz Zilyas'ın yüksek lisans tezinden türetilmiştir.

Yazar Katkıları

Tüm yazarlar usulüne uygun olarak makaleye katkıda bulunmuştur. Bu makale yüksek lisans tezine dayanmaktadır. Yazarlar yüksek lisans tez öğrencisi ve tez danışmanıdır.

-Çalışma kavramı ve dizaynı (Z, Y)
-Veri elde etme (Z)
-Veri çözümleme ve tefsiri (Z, Y)
-Taslağın oluşturulması (Z)
-Revizyon (Y)

.

Kaynakça

[1] Çelenk, Süleyman. 2003. "Okul Başarısının Ön Koşulu: Okul Aile Dayanışması", *Çelenk, S. İlköğretim-Online 2 (2), 2003 sf. 28-34*

[2] Çelenk, Süleyman. 2003. "Okul Başarısının Ön Koşulu: Okul Aile Dayanışması", *Çelenk, S. İlköğretim-Online 2 (2), 2003 sf. 28-34*

[3] Yavuz, Erol. "Konutlarda Enerji Tüketimi Kestirimi İçin Derin Öğrenme Ve Makine Öğrenme Yöntemlerinin Karşılaştırılması", 2020,İstanbul,*Yüksek Lisans Tezi*

[4] Gök, Murat. 2017. "Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi",*Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA*

[5] Akşehir, Zinnet Duygu ve Kılıç, Erdal. 2019 "Makine Öğrenmesi Teknikleri ile Banka Hisse Senetlerinin Fiyat Tahmini", *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi (2019 Cilt:12- Sayı:2)*

[6] Onur Sevli ve Ali Tezcan Sarızeybek. 2019. "Makine Öğrenmesi Yöntemleri İle Banka Müşterilerinin Kredi Alma Eğiliminin Araştırma Analizi." , *Zeki Sistemler Teori ve Uygulamaları Dergisi 5(2) 2022 137-144*

[7] Yağcı, Mustafa. 2022. "Eğitsel Veri Madenciliği: Makine Öğrenimi Algoritmalarını Kullanarak Öğrencilerin Akademik Performansının Tahmini", *Smart Learning Environments 9:11 2022, Kırşehir*

[8] Cruz-Jesus, F., Castelli M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M.,&Rosa-Louro, A. 2020 " *Using artifcial intelligence methods to assess academic achievement in public high schools of a European Union country."* Heliyon.

[9] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). *"Educational data mining : Predictive analysis of academic performance of public school students in the capital of Brazil*" Journal of Business Research, 335–343

[10] Xu, X., Wang, J., Peng, H., & Wu, R. (2019)*." Prediction of academic performance associated with internet usage behaviors using machine learning algorithms."* 166–173.

[11] Hofait, A., & Schyns, M. (2017*)." Early detection of university students with potential difculties"* Decision Support System 1–11

[12] Chui, K.; Fung, D.; Lytras, M.; Lam, T. 2020, *"Predicting at-risk university students in a virtual learning environment via a machine learning algorithm"*,107

[13] Nieto, Y.; García-Díaz, V.; Montenegro, C.; Crespo, R.G. 2019, *"Supporting academic decision making at higher educational institutions using machine learning-based algorithms"*,4145–4153

[14] Ahmad, Z., & Shahzadi, E. (2018) *"Prediction of students' academic performance using artifcial neural network"*, Bulletin of Education and Research,40(3)

[15] Dam, Hasan . 2008. "Öğrencinin Okul Başarısında Aile Faktörü",Hitit Üniversitesi İlahiyat Fakültesi Dergisi, 2008/2, c. 7, sayı: 14, ss. 75-99.

[16] Aslanargun, Engin, Bozkurt, Sinan, Sarıoğlu, Selma. 2016 "*Sosyo Ekonomik Değişkenlerin Öğrencilerin Akademik Başarısı Üzerine Etkileri*",Uşak Üniversitesi Sosyal Bilimler Dergisi 9/3

[17] Nilsson ,Nils J. 1998. "*Introduction To Machine Learning*", Stanford University

[18] Chollet, François. 2019. "Deep Learning With Python", *Buzdağı Yayınevi, p.5*

[19] Ayık Ziya, Özdemir Abdulkadir, Yavuz Uğur. 2007. "Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi*", Atatürk Üniversitesi Sosyal Bilimler Enstitsü Dergisi. 10(2): 441-454*

[20] *Ulgen,Kaan .2017,* "Makine Öğrenimi Bölüm-2(K-NN)*", https://l24.im/1kg4*

[21] Breiman,Leo. 2001. *" Random Forest"*, University Of California Berkeley,CA 94720, *p.2*

[22] Grubbs, 1969. "Procedures for Detecting Outlying Observations in Samples. Technometrics"